REGION-BASED MULTIMODAL IMAGE FUSION USING ICA BASES

Nedeljko Cvejic, John Lewis, David Bull, Nishan Canagarajah

Department of Electrical and Electronic Engineering
University of Bristol
Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, United Kingdom

ABSTRACT

In this paper, we present a novel region-based multimodal image fusion algorithm in the ICA domain. It uses segmentation to determine the most important regions in the input images and consequently fuses the ICA coefficients from the given regions. The proposed method exhibits considerably higher performance than the basic ICA algorithm and shows improvement over other state-of-the-art algorithms.

Index Terms— Image fusion, joint segmentation, region-based fusion, Independent component analysis

1. INTRODUCTION

Rapid advances in the areas of sensor technology and communication networks have lead to a need for processing that can efficiently fuse information from different sensors into a single composite signal. Image and video fusion is a subarea of the more general topic of data fusion, dealing with image and video data [1]. Multi-sensor data often presents complementary information about a scene or object of interest, and thus image fusion provides an effective method for comparison and analysis of such data. There are several benefits of multi-sensor image fusion: wider spatial and temporal coverage, extended range of operation, increased robustness of the system performance and enhanced detection and classification capabilities.

The image fusion process can be performed at different levels of information representation: signal, pixel, feature and symbolic level. Feature-level fusion methods include region-based imaged fusion. Here images to be fused are initially segmented into a set of distinctive regions. Various properties of the regions obtained by segmentation can be used to determine which features from which images are to be included in the fused image. This has advantages over pixel-based methods as more intelligent semantic fusion rules can be considered based on actual features in the image, rather than on single or arbitrary groups of pixels.

Nikolov et al [1] proposed a classification of image fusion algorithms into spatial domain and transform domain techniques. Instead of using a standard bases system, such as the DFT, the mother wavelet or cosine bases of the DCT, one can train a set of bases that are suitable for a specific type of images. A training set of image patches, which are acquired randomly from images of similar content, can be used to train a set of statistically independent bases. This is known as Independent Component Analysis (ICA) [2]. Recently, several algorithms have been proposed [3, 4], in which ICA bases are used for transform domain image fusion. In this paper, we refine the approach using a novel multimodal image fusion algorithm in the ICA domain. Segmentation is used to determine the most important regions in the input images and consequently the ICA coefficients are used to fuse the given regions.

This work has been funded by the UK Ministry of Defence Data and Information Fusion Defence Technology Centre.

2. BACKGROUND REVIEW

In order to obtain a set of statistically independent bases for image fusion in the ICA domain, training is performed with a predefined set of images. Training images are selected in such a way that the content and statistical properties are similar for the training images and the images to be fused. An input image $i(x, y)$ is randomly windowed using a rectangular window w of size $N \times N$. The result of windowing is an "image patch" which is defined as [3]:

$$p(m, n) = w \cdot i(m_0 - N/2 + m, n_0 - N/2 + n) \qquad (1)$$

where m and n take integer values from the interval $[0, N - 1]$. Each image patch $p(m, n)$ can be represented by a linear combination of a set of M basis patches $b_i(m, n)$:

$$p(m, n) = \sum_{i=1}^{M} v_i b_i(m, n) \qquad (2)$$

where $v_1, v_2, ..., v_M$ stand for the projections of the original image patch on the basis patch, i.e. $v_i = \langle p(m, n), b_i(m, n) \rangle$. A 2D representation of the image patches can be simplified to a 1D representation, using lexicographic ordering. This implies that an image patch $p(m, n)$ is reshaped into a vector $\underline{p}$, mapping all the elements from the image patch matrix to the vector in a row-wise fashion. Decomposition of image

patches into a linear combination of basis patches can then be expressed as follows:

$$\underline{p}(t) = \sum_{i=1}^{M} v_i(t)\underline{b}_i = [\underline{b}_1\underline{b}_2...\underline{b}_M] \cdot \begin{bmatrix} v_1(t) \\ v_2(t) \\ ... \\ v_M(t) \end{bmatrix} \quad (3)$$

where t represents the image patch index. If we denote $B = [b_1 b_2 ... b_M]$ and $v(t) = [v_1 v_2 ... v_M]^T$, then equation (3) reduces to:

$$\underline{p}(t) = B\underline{v}(t) \quad (4)$$

$$\underline{v}(t) = B^{-1}\underline{p}(t) = A\underline{p}(t) \quad (5)$$

Thus, $B = [b_1 b_2 ... b_M]^T$ represents an unknown mixing matrix (analysis kernel) and $A = [a_1 a_2 ... a_M]^T$ the unmixing matrix (synthesis kernel). This transform projects the observed signal $\underline{p}(t)$ on a set of basis vectors. The aim is to estimate a finite set of $K < N^2$ basis vectors that will be capable of capturing most of the input image properties and structure.

In the first stage of basis estimation the Principal Component Analysis (PCA) is used for dimensionality reduction. This is obtained by eigenvalue decomposition of the data correlation matrix $C = E\{\underline{p}\,\underline{p}^T\}$. The eigenvalues of the correlation matrix illustrate the significance of their corresponding basis vector. If V is the obtained $K \times N$ PCA matrix, the input image patches are transformed by:

$$\underline{z}(t) = V\underline{p}(t) \quad (6)$$

After the PCA preprocessing step we select the statistically independent basis vectors using the optimisation of the negentropy. The following rule defines a FastICA approach that optimises negentropy, as proposed in [2]:

$$\underline{a}_i^+ \leftarrow \varepsilon\{\underline{a}_i \phi(\underline{a}_i^T \underline{z})\} - \varepsilon\{\phi'(\underline{a}_i^T \underline{z})\}\underline{a}_i \qquad 1 \le i \le K \quad (7)$$

$$A \leftarrow A(A^T A)^{-0.5} \quad (8)$$

where $\phi(x) = -\partial G(x)/\partial x$ defines the statistical properties $G(x) = \log p(x)$ of the signals in the transform domain [2]. In our implementation we used:

$$G(x) = \alpha\sqrt{\varepsilon + x} + \beta \quad (9)$$

where α and β are constants and ε is a small constant to tackle numerical instability, in the case that $x \to 0$ [2].

After the input image patches $\underline{p}(t)$ are transformed to their ICA domain representations $\underline{v}_k(t)$, we can perform image fusion in the ICA domain in the same manner as it is performed in e.g. the wavelet domain. The equivalent vectors $\underline{v}_k(t)$ from each image are combined in the ICA domain to obtain a new image $\underline{v}_f(t)$. The method that combines the coefficients in the ICA domain is called the "fusion rule". After the composite image $v_f(t)$ is constructed in the ICA domain, we can move back to the spatial domain, using the synthesis kernel A, and synthesise the image $i_f(x, y)$. Several features can be employed in the estimation of the contribution of each input image to the fused output image. In [3], the authors proposed the mean absolute value of each $N \times N$patch in the transform domain, as an activity indicator:

$$E_k(t) = \|\underline{v}_k(t)\|, \qquad k = 1, ..., T \quad (10)$$

As the ICA bases tend to focus on edge information, large values for $E_k(t)$ correspond to increased activity in the frame, e.g. the existence of edges. Based on this observation, the authors in [3] divide the transform domain patches in two groups. The first group consists of the regions that contain details ($E_k(t)$ larger then a threshold) and the second group contains the region with background information ($E_k(t)$ smaller then a threshold). The threshold that determines whether a region is "active" or "non-active" is set heuristically. As a result, the segmentation map $s_k(t)$ is created for each input image [3]:

$$s_k(t) = \begin{cases} 1 & \text{if } E_k(t) > \frac{2}{T}\sum_{k=1}^{T} E_k(t) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

The segmentation maps of input images are combined to form a single segmentation map, using the logical OR operator [3]:

$$s(t) = OR\{s_1(t), s_2(t), ..., s_T(t)\} \quad (12)$$

After the input images are segmented into active and non-active regions, two different fusion rules are used for fusion of each group of regions [3]. Namely, active regions are fused using the "max-abs" rule, while non-active regions are fused using the "mean" rule. The "max-abs" rule fuses two input coefficients/vectors by selecting the one with higher absolute value. In the "mean" fusion rule the fused coefficient/vector is equal to the mean value of the two input coefficients/vectors.

3. PROPOSED METHOD

In this paper we focus on the fusion of infra-red (IR) and visible images, although methods can be generalized to other modalities. Because the threshold that determines the "activity" of a region is set heuristically, the regions obtained by thresholding of the ICA coefficients do not correspond always to objects in the images to be fused. Our experiments showed that important objects in the IR input images (e.g. a person or a smaller object) are often masked by textured high-energy background in the visual image. In this case the important objects from the IR image become blurred or, in extreme cases, completely masked. Therefore, we perform segmentation in the spatial domain and then fuse patches from separate regions separately. This differs from the methods in [3, 4] where the fusion was performed on a more general, pixel level.

3.1. The segmentation algorithm

The quality of the segmentation algorithm is of vital importance to the fusion process. An adapted version of the combined morphological−spectral unsupervised image segmentation algorithm is used, which is described in [5], enabling it to handle multi-modal images. The algorithm works in two stages. The first stage produces an initial segmentation by using both textured and non-textured regions. The detail coefficients of the DT-CWT are used to process texture. The gradient function is applied to all levels and orientations of the DT-CWT coefficients and up-sampled to be combined with the gradient of the intensity information to give a perceptual gradient. The larger gradients indicate possible edge locations. The watershed transform of the perceptual gradient gives an initial segmentation. The second stage uses these primitive regions to produce a graph representation of the image which is processed using a spectral clustering technique.

The method can use either intensity information or textural information or both to obtain the segmentation map. This flexibility is useful for multi-modal fusion where some a priori information of the sensor types is known. For example, IR images tend to lack textural information with most features having a similar intensity value throughout the region. Therefore, we used an intensity only segmentation map, as it gives better results than a texture based segmentation.

The segmentation can be performed either separately or jointly. For separate segmentation, each of the input images generates an independent segmentation map for each image.

$$S_1 = \sigma(i_1, D_1), \ldots, S_N = \sigma(i_N, D_N) \tag{13}$$

where D_n represent detail coefficients of the DT-CWT used in segmentation. Alternatively, information from all images could be used to produce a joint segmentation map.

$$S_{joint} = \sigma(i_1 \cdots i_N, D_1 \cdots D_N) \tag{14}$$

In general, jointly segmented images work better for fusion [6]. This is because the segmentation map will contain a minimum number of regions to represent all the features in the scene most efficiently. A problem can occur for separately segmented images, where different images have different features or features which appear as slightly different sizes in different modalities. Where regions partially overlap, if the overlapped region is incorrectly dealt with, artefacts will be introduced and the extra regions created to deal with the overlap will increase the time taken to fuse the images.

3.2. Calculation of priority and fusion rules

After the images are jointly segmented it is essential to determine the importance of regions in each of the input images. We have decided to use the normalized Shannon entropy of a region as the priority. Thus, the priority $P(r_{t_n})$ is given as:

$$P(r_{t_n}) = \frac{1}{|r_{t_n}|} \sum_{\forall\theta,\forall l,(x,y)\in r_{t_n}} d^2_{n(\theta,l)}(x,y) \log d^2_{n(\theta,l)}(x,y) \tag{15}$$

with the convention $0\log(0) = 0$, where $|r_{t_n}|$ is the size of the region r_{t_n} in input image n and $d_{n(\theta,l)}(x,y) \in D_{n(\theta,l)}$ detail coefficients of the DT-CWT used in segmentation. Finally, a mask M is generated that determines which image each region should come from in the fused image. An example of the IR input image, visual input image, performed joint segmentation and the image fusion mask is given in Fig. 1.

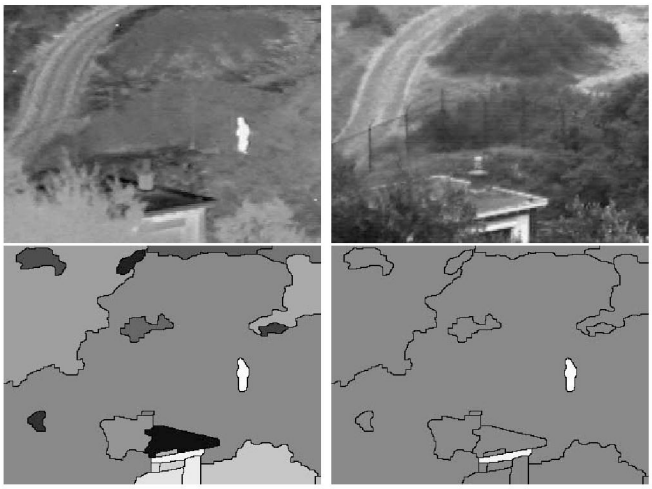

Fig. 1. Segmentation and region selection prior to fusion. Top: IR input image (left), visible input image (right). Bottom: Regions obtained by joint segmentation of input images (left), image mask: white from IR, grey from visible (right).

4. EXPERIMENTAL RESULTS

The proposed image fusion method was tested in the multi-modal scenario with two input images: infrared and visible. In order to make a comparison between the proposed method and the standard ICA method, the images were fused using the approach described in [3]. We compared these results with a simple averaging method, the ratio pyramid method, the Laplace transform (LT) and the dual-tree complex wavelet transform (DT-CWT)[6]. In the multiresolution methods (LT, DT-CWT) a 5-level decomposition is used and fusion is performed by selecting the coefficient with a maximum absolute value, except for the case of the lowest resolution subband where the mean value is used. Before performing image fusion, the ICA bases were trained using a set of ten images with content comparable to the test set. The number of rectangular patches ($N = 8$) used for training was 10000, randomly selected from the training set. The lexicographic ordering was applied to the image patches and then PCA performed. Fol-

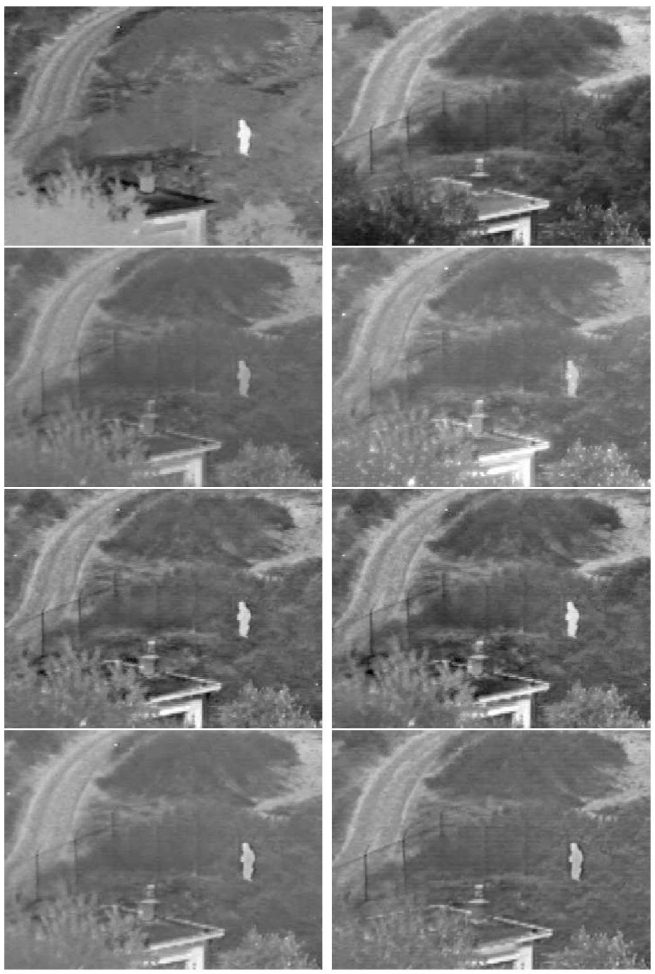

Fig. 2. Top: input IR image (left) and input visible image (right); Second row: fused image using averaging (left) and ratio pyramid (right); Third row: fused image using DT-CWT (left) and LT (right); Bottom row: fused image using standard ICA (left) and region-based ICA method (right)

lowing this, the 32 most important bases ($K = 32$) were selected, according to the eigenvalues corresponding to these bases. After that, the ICA update rule in (7) was iterated for $L = 3$ (3x3 neighbourhood) until convergence.

ICA coefficients are combined using the principle described in Section 2 for comparison. The images to be fused were then segmented, regions and image masks determined for each of them and then ICA fusion performed on these regions using the "max-abs" fusion rule. Example input images and fused outputs are given in Fig. 2. It is clear that the fused image obtained using the proposed algorithm incorporate more detail from the visible image together with the important objects from the IR image, compared to the standard ICA method. The data presented in Table 1 confirms this conclusion, using both the Petrovic [7] and the Piella metric [8]. The proposed method exhibits considerably higher performance than the basic ICA algorithm and improvement over other state-of-the-art algorithms.

Table 1. Performance of the image fusion methods measured by standard fusion metrics.

Metric	Method	UN 1812	Trees 4917	Octec 22
Piella	Average	0.866	0.962	0.872
	Laplace	0.914	0.969	0.939
	DT-CWT	0.912	0.969	**0.941**
	Ratio	0.862	0.960	0.876
	ICA	0.872	0.962	0.889
	R-B ICA	**0.921**	**0.974**	0.940
Petrovic	Average	0.347	0.513	0.436
	Laplace	0.501	0.599	0.767
	DT-CWT	0.462	0.600	0.768
	Ratio	0.413	0.533	0.503
	ICA	0.415	0.539	0.613
	R-B ICA	**0.548**	**0.636**	**0.784**

5. REFERENCES

[1] R. Blum and Z. Liu, *Multi-sensor Image Fusion And Its Applications*, CRC Press, London, UK, 2005.

[2] A. Hyvrinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, London, UK, 2001.

[3] N. Mitianoudis and T. Stathaki, "Pixel-based and region-based image fusion schemes using ICA bases," *Information Fusion*, to appear.

[4] N. Cvejic, D. Bull, and N. Canagarajah, "A novel ICA domain multimodal image fusion algorithm," in *Proc. SPIE Defense and Security Symposium*. to appear, Orlando, FL.

[5] R. O'Callaghan and D. Bull, "Combined morphological-spectral unsupervised image segmentation," *IEEE Transactions on Image Processing*, vol. 14, pp. 49–62, 2005.

[6] J. Lewis, R. O'Callaghan, S. Nikolov, D. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Information Fusion*, to appear.

[7] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE International Conference on Image Processing*. 2003, pp. 173–176, Barcelona, Spain.

[8] C. Xydeas and V. Petrovic, "Objective pixel-level image fusion performance measure," in *Proc. SPIE*. 2000, pp. 88–99, Orlando, FL.